AD-A274 212

TION PAGE

(2)

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
FINAL/01 JUL 92 TO 30 JUN 93

4. TITLE AND SUBTITLE
PARAMETRIC TIME-SCALE METHODS IN
SIGNAL ANALYSIS (U)

5. FUNDING NUMBERS
2304/ES
F49620-92-J-0378

6. AUTHOR(S)
Professor R. Kumaresan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Rhode Island
Dept of Electrical Engineering
Kingston, RI 02881-0805

8. PERFORMING ORGANIZATION REPORT NUMBER
AEOSR-TR- 93 0002

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
110 DUNCAN AVE, SUITE B115
BOLLING AFB DC 20332-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
F49620-92-J-0378

DTIC
ELECTE
DEC 30 1993
S
A

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

12b. DISTRIBUTION CODE
UL

13. ABSTRACT (Maximum 200 words)

The researchers have separated the Kaiser-Teager algorithm for separating the contributions of the amplitude modulation and frequency modulation of a signal consisting of multiple components. To achieve this separation, they use an instantaneous non-linear operator, which turns out to be the determinant of a Toeplitz matrix formed with the signal samples. Because of its instantaneously adaptive nature, they can use this algorithm to track parameter variations are not too rapid. The researchers demonstrate this using a synthetic signal containing two AM-FM signal components and a speech signal. They also point out the methods relationship to Prony's method. The researchers proposed a demodulator-filter approach for tracking harmonic signals and applied it to speech processing.

14. SUBJECT TERMS

15. NUMBER OF PAGES
8

16. PRICE CODE

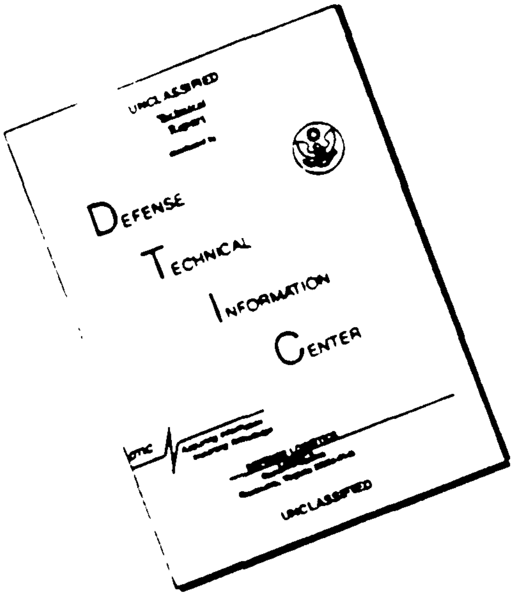
17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
SAR(SAME AS REPORT)

DISCLAIMER NOTICE



UNCLASSIFIED

DEFENSE
TECHNICAL
INFORMATION
CENTER

DTIC

UNCLASSIFIED

FINAL REPORT

FOR

UNIVERSITY OF RHODE ISLAND

F49620-92-J-0378

1 JUL 92 - 30 JUN 93

Final report for the AFOSR Contract F-4620-92-J-0378

Under this contract we addressed two topics. Firstly, we proposed a non-linear operator for tracking multi-component-signal parameter. This topic is described in paper #1. Secondly we proposed a demodulator-filter approach for tracking harmonic signals and applied it to speech processing. This is described in paper #2.

Paper #1

INSTANTANEOUS NON-LINEAR OPERATORS FOR TRACKING MULTICOMPONENT SIGNAL PARAMETERS[*]

R. Kumaresan[1] *A. G. Sadasiv*[1] *C. S. Ramalingam*[1] *J. F. Kaiser*[2]

[1]Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881.
[2]Department of Electrical Engineering, Rutgers University, Piscataway, NJ 08855.

ABSTRACT

We have extended the Kaiser–Teager algorithm for separating the contributions of the amplitude modulation and frequency modulation of a single sinusoid to a signal consisting of multiple components. To achieve this separation, we use an instantaneous non-linear operator, which turns out to be the determinant of a Toeplitz matrix formed with the signal samples. Because of its instantaneously adaptive nature, we can use this algorithm to track parameter variations in the signal components, provided these variations are not too rapid. We demonstrate this using a synthetic signal containing two AM-FM signal components and a speech signal. We also point out the method's relationship to Prony's method.

1. INTRODUCTION

In many applications, such as speech processing [1, 2], radar signal processing [3], one can model the observed signal as a linear combination of a small number of sinusoidal signals that are slowly time-varying in both amplitude and frequency. In such cases, a signal consisting of m components may be parametrized as

$$x(t) = \sum_{k=1}^{m} A_k(t)\cos(\theta_k(t)), \tag{1}$$

where $\frac{d\theta_k(t)}{dt} = \omega_k + \phi_k(t)$. ω_k is the nominal carrier frequency of the k^{th} component, while $\phi_k(t)$ and $A_k(t)$ are the time-varying frequency and amplitude, respectively. Given the signal $x(t)$, one wishes to determine and track the amplitude envelopes and the frequency trajectories of the individual signal components. Traditional short-time Fourier transform methods are not always successful in processing such signals due to their limited resolution. The so-called time-frequency representations, such as the Wigner distribution, have their own problems with multicomponent signals [4].

Recently, Kaiser [5, 6] introduced a novel approach to track the frequency and amplitude variations of a single component signal by applying what is called an energy operator. Originally devised by Teager, the discrete-time version of the energy operator $\Psi(\cdot)$ is an instantaneous non-linear function of the samples of a signal:

$$\Psi(x_n) = x_n^2 - x_{n-1}x_{n+1}. \tag{2}$$

If x_n represents a sinewave, *i.e.*, $x_n = A_1\cos(\omega_1 n + \phi_1)$, whose radian frequency ω_1 is constant, then the 'energy' $\Psi(x_n) = A_1^2\sin^2\omega_1$ [5]. Observe that this quantity is time-invariant. Using this functional dependence of energy on the frequency and amplitude of a sinusoidal signal, Kaiser and his collaborators [7] devised methods to separate the contribution of the amplitude and frequency of a signal. They then applied this method to approximately determining the amplitude and frequency variations of an amplitude/frequency modulated (AM-FM) signal. Although their method is computationally simple and instantaneously adaptive in nature, it cannot be directly applied to tracking amplitude and frequency variations of signals composed of multiple components as in (1). For such multicomponent signals, Kaiser [5] advocates first separating the signals into frequency components by filtering.

In this paper we extend the method in [7] to simultaneously tracking the amplitude and frequency variations of multicomponent signals. This is accomplished by using certain instantaneous non-linear operators on the signal samples. These can be motivated by viewing the Teager energy operator $\Psi(x_n)$ as the determinant of a matrix:

$$\Psi(x_n) = \mathrm{Det}\begin{bmatrix} x_n & x_{n+1} \\ x_{n-1} & x_n \end{bmatrix}. \tag{3}$$

As mentioned before, this determinant is time-invariant for a single sinusoidal signal with constant frequency. We have shown [8] that this time-invariance property of the determinant can be extended to an appropriately constructed higher-order $m \times m$ matrix, whose elements are the samples of a sum of m sinusoids. A similar result has been obtained for continuous-time signals as well [8]. In section 2 we have outlined this result. In section 3 we show how these results can be related to the energy separation algorithms derived by Maragos, Quatieri, and Kaiser [7]. Further, in section 4, using the functional dependence of these determinants on the frequencies and amplitudes of the sinewave components, we describe the two-component AM-FM separation algorithm. We obtain explicit expressions for the instantaneous frequencies of two (real-valued) component signals. This can be extended to up to four components

---

[*] *THIS WORK WAS SUPPORTED BY AN AIR FORCE CONTRACT AFOSR 49620-92-J-0378*

ily. However, beyond four components one has to root polynomial to obtain the frequencies. This is reminiscent the well-known Prony's method [9].

2. TIME INVARIANCE OF THE DETERMINANT OF A TOEPLITZ MATRIX

Let x_n be the samples of a linear combination of m omplex-valued, discrete-time sinusoidal signals with arbitrary phases and non-zero amplitudes, *i.e.*,

$$x_n = \sum_{k=1}^{m} A_k e^{j(\omega_k n + \phi_k)} \quad . \tag{4}$$

The distinct radian frequencies $\omega_k \in [0, 2\pi)$ have no particular relationship between them. Consider the following $m \times m$ Toeplitz matrix $\mathbf{X}_m(n)$ formed with the elements of the sequence x_n:

$$\mathbf{X}_m(n) = \begin{bmatrix} x_n & x_{n+1} & \cdots & x_{n+m-1} \\ x_{n-1} & x_n & \cdots & x_{n+m-2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-m+1} & x_{n-m+2} & \cdots & x_n \end{bmatrix} . \tag{5}$$

Using the definition of x_n in (4), we can decompose $\mathbf{X}_m(n)$ into a product of three matrices to facilitate the computation of its determinant, $\Delta\mathbf{x}$, where the dependence on m and n is not explicitly shown for notational simplicity. This decomposition may be written as

$$\mathbf{X}_m(n) = \mathbf{V}\Lambda(n)\mathbf{V}^H . \tag{6}$$

The superscript 'H' denotes Hermitian transpose. The diagonal matrix $\Lambda(n)$ has $\lambda_k = A_k e^{j(\omega_k n + \phi_k)}$, $k = 1, 2, \ldots, m$ as its diagonal entries. $\mathbf{V}$ is a Vandermonde matrix [10] with entries $v_{kl} = e^{j\omega_l (k-1)}$, $k, l = 1, 2, \ldots, m$. Note that the diagonal matrix in the middle is the one that shows dependence on the time index. The determinant of $\mathbf{V}$ is $\prod_{k,l=1,k<l}^{m} (e^{j\omega_k} - e^{j\omega_l})$ [10] and the determinant of $\Lambda(n)$ is $\prod_{k=1}^{m} A_k e^{j(\omega_k n + \phi_k)}$. Hence, after some simplification,

$$\Delta\mathbf{x} = \prod_{k=1}^{m} A_k e^{j(\omega_k n + \phi_k)} \prod_{\substack{k,l=1 \\ k<l}}^{m} 4 \sin^2 \left(\frac{\omega_k - \omega_l}{2} \right) . \tag{7}$$

Since the complex-valued factor in the above formula has unit magnitude, we observe that the magnitude of $\Delta\mathbf{x}$ is invariant with the time index.

If the signal is composed of $m/2$ real-valued sinusoids, i.e., $x_n = \sum_{k=1}^{m/2} A_k \cos(\omega_k n + \phi_k)$, then $\omega_{k+m/2} = -\omega_k$ and $\phi_{k+m/2} = -\phi_k$ for $k = 1, 2, \ldots, m/2$. This leads to further simplification of (7) for $\Delta\mathbf{x}$:

$$\begin{aligned} & \prod_{k=1}^{m/2} A_k^2 \sin^2 \omega_k \\ & \quad \prod_{\substack{k,l=1 \\ k<l}}^{m/2} \left(4 \sin^2 \left(\frac{\omega_k - \omega_l}{2} \right) \right)^2 \left(4 \sin^2 \left(\frac{\omega_k + \omega_l}{2} \right) \right)^2 . \end{aligned} \tag{8}$$

Note that this determinant is always positive, and of course, time-invariant. The essence of the above results can be expressed in words as follows: For a sequence x_n, composed of a sum of m complex-valued sinewaves with arbitrary frequencies, phases, and amplitudes, there exists an instantaneous non-linear function involving $(2m - 1)$ samples that takes on a constant magnitude anywhere on the time-index axis. A similar statement is true for real-valued sinewaves. This appears to be a useful fact as demonstrated below. Similar results can also be obtained for continuous-time signals [8].

Now assume that the sequence x_n in (4) is passed through a linear time-invariant filter with impulse response $h_n \xrightarrow{\text{DTFT}} H(e^{j\omega})$. Then the output sequence y_n may be written as

$$y_n = \sum_{k=1}^{m} A'_k e^{j(\omega_k n + \phi_k)} \quad , \tag{9}$$

where $A'_k = A_k H(e^{j\omega_k})$. This result follows from the fact that complex exponentials are eigenfunctions for a linear time-invariant filter. If we construct an $m \times m$ Toeplitz matrix $\mathbf{Y}_m(n)$ (similar to $\mathbf{X}_m(n)$ in (5)) with the elements of the sequence y_n, since the expression for y_n is akin to that of x_n in (4), we can write down the determinant of $\mathbf{Y}_m(n)$ using (7) as

$$\Delta\mathbf{y} = \prod_{k=1}^{m} A'_k e^{j(\omega_k n + \phi_k)} \prod_{\substack{k,l=1 \\ k<l}}^{m} 4 \sin^2 \left(\frac{\omega_k - \omega_l}{2} \right) . \tag{10}$$

Therefore, the determinants of $\mathbf{Y}_m(n)$ and $\mathbf{X}_m(n)$ are related as follows:

$$\Delta\mathbf{y} = \Delta\mathbf{x} \cdot \prod_{k=1}^{m} H(e^{j\omega_k}) \quad . \tag{11}$$

Next, we show the relationship of the above results to the Discrete Energy Separation Algorithms (DESA) of Maragos *et al.* [7].

3. RELATION TO THE DESA ALGORITHMS

In [7] Maragos *et al.* have proposed two algorithms, *viz.*, DESA-2 and DESA-1, to separate amplitude modulation from frequency modulation using the energy operator. Using the results derived in the previous section we first show how we can obtain their algorithms. Extension to more than one component is discussed in the next two sections.

Let $x_n = A_1(n)\cos(\omega_1(n) + \phi_1)$ correspond to a single AM-FM signal. Let us assume that $A_1(n)$ and $\omega_1(n)$ are only slowly varying and that they can be approximated as constants over any consecutive three-sample period. Since x_n consists of two complex exponentials $(m = 2)$, from (8) $\Delta\mathbf{x} \approx A_1^2(n) \sin^2[\omega_1(n)]$. Next, filter x_n through a filter with impulse response $h_n = \{\frac{1}{2}, 0, -\frac{1}{2}\} \xrightarrow{\text{DTFT}} \frac{1}{2}(e^{j\omega} - e^{-j\omega})$ to yield y_n. Using (11), $\Delta\mathbf{y}$ is found to be

$$\begin{aligned} \Delta\mathbf{y} &\approx \Delta\mathbf{x}\, H(e^{j\omega_1(n)})\, H(e^{-j\omega_1(n)}) \\ &= A_1^2(n) \sin^2[\omega_1(n)] \, \frac{1}{4} \left| e^{j\omega_1(n)} - e^{-j\omega_1(n)} \right|^2 \\ &= A_1^2(n) \sin^4[\omega_1(n)] \quad . \end{aligned} \tag{12}$$

which immediately leads to DESA-2. To obtain DESA-1 we observe that y_n is the output of a filter whose transfer function is $H_1(e^{j\omega}) = 1 - e^{-j\omega}$, while z_n is the result of filtering x_n by $H_2(e^{j\omega}) = e^{j\omega} - 1$, and applying (11) yields Eq. (5) in [7]. Thus, these two of the algorithms proposed in [7] can be obtained as special cases of the above expressions.

The algorithms devised in [7] have the disadvantage of being applicable to only single-component AM-FM signals. Using (11) we next show how to separate amplitude from frequency modulation in the case of a two-component AM-FM signal.

4. A TWO-COMPONENT AM-FM SIGNAL SEPARATION ALGORITHM

We now extend the above algorithm to separate a (real-valued) two-component AM-FM signal. Let $x_n = A_1(n)\cos(\omega_1(n) + \phi_1) + A_2(n)\cos(\omega_2(n) + \phi_2)$. For this case, $m = 4$. Next, we filter x_n by $H_1(e^{j\omega}) = 1 + e^{-j\omega}$ to get $y_n^{(1)}$ and by $H_2(e^{j\omega}) = 1 - e^{-j\omega}$ to get $y_n^{(2)}$. As before, we assume that $A_i(n)$ and $\omega_i(n)$, $i = 1, 2$ are only slowly varying so that they can be approximated as constants over any period of $2m = 8$ samples. Hence, as before, using (11)

$$\begin{aligned} \frac{\Delta \mathbf{y}_1}{\Delta \mathbf{x}} &\approx \left|1 + e^{-j\omega_1(n)}\right|^2 \left|1 + e^{-j\omega_2(n)}\right|^2 , \\ \frac{\Delta \mathbf{y}_2}{\Delta \mathbf{x}} &\approx \left|1 - e^{-j\omega_1(n)}\right|^2 \left|1 - e^{-j\omega_2(n)}\right|^2 . \end{aligned}$$

Simplifying, we obtain

$$\frac{1}{2}\sqrt{\frac{\Delta \mathbf{y}_1}{\Delta \mathbf{x}}} = \cos\frac{\omega_1(n) - \omega_2(n)}{2} + \cos\frac{\omega_1(n) + \omega_2(n)}{2}, \quad (13)$$

$$\frac{1}{2}\sqrt{\frac{\Delta \mathbf{y}_2}{\Delta \mathbf{x}}} = \cos\frac{\omega_1(n) - \omega_2(n)}{2} - \cos\frac{\omega_1(n) + \omega_2(n)}{2}. \quad (14)$$

From the above two equations we can solve for $\omega_1(n)$ and $\omega_2(n)$, respectively. Once the frequencies have been computed, the amplitudes can be obtained by solving a set of linear equations in the least-squares sense.

5. MULTICOMPONENT AM-FM SEPARATION ALGORITHM

The AM-FM separation algorithm developed in the previous section can be generalized to deal with an m-component signal. To this end, consider m distinct length-two filters, with transfer functions of the form $H_k(e^{j\omega}) = 1 + a_k e^{j\omega}$, $1 \le k \le m$. The a_k are, in general, complex. Denote by $y_n^{(k)}$ the output of the filter $H_k(e^{j\omega})$ with x_n as its input. Then, applying (11), the determinants of the $m \times m$ Toeplitz matrices formed from the samples of x_n and $y_n^{(k)}$ are related by

$$\Delta \mathbf{y}_k \approx \Delta \mathbf{x} \cdot \prod_{l=1}^{m} \left(1 + a_k e^{j\omega_l(n)}\right) \quad (15)$$

for $1 \le k \le m$. Or, using matrix notation,

$$\begin{pmatrix} a_1 & a_1^2 & \cdots & a_1^m \\ a_2 & a_2^2 & \cdots & a_2^m \\ \vdots & \vdots & \vdots & \vdots \\ a_m & a_m^2 & \cdots & a_m^m \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \approx \begin{pmatrix} \frac{\Delta \mathbf{y}_1}{\Delta \mathbf{x}} - 1 \\ \vdots \\ \frac{\Delta \mathbf{y}_m}{\Delta \mathbf{x}} - 1 \end{pmatrix}, \quad (16)$$

where b_k is the $(k+1)^{st}$ term in the expansion of $\prod_{l=1}^{m}\left(1 + e^{j\omega_l(n)}\right)$. Since the a_k's are distinct, the Vandermonde matrix on the left-hand side is non-singular. Hence (16) can be solved by matrix inversion. The frequencies $\omega_k(n)$ are easily seen to be the angles of the roots of the polynomial $1 + \sum_{l=1}^{m} b_l z^{-l}$. Once again, the assumption is that the parameters of the AM-FM signal can be approximated as constants over any $2m$-sample interval.

6. SIMULATIONS AND DISCUSSION

As an illustrative example of the above techniques, consider the following two-component AM-FM signal of the form $s(n) = A_1(n)s_1(n) + A_2(n)s_2(n)$, where

$$s_1(n) = \begin{cases} \cos\left(0.25\pi n + \frac{\pi n^2}{8000}\right) & n = 0, 1, \ldots, 200 \\ \cos\left(0.35\pi n - \frac{\pi n^2}{8000}\right) & n = 201, 202, \ldots, 399 \end{cases}$$

$$s_2(n) = \begin{cases} \cos\left(0.5\pi n - \frac{\pi n^2}{8000}\right) & n = 0, 1, \ldots, 199 \\ \cos\left(0.4\pi n + \frac{\pi n^2}{8000}\right) & n = 200, 201, \ldots, 399 \end{cases}$$

$$\left.\begin{aligned} A_1(n) &= 1 - 0.25\cos\left(\frac{\pi n}{200} + \frac{\pi}{3}\right) \\ A_2(n) &= 1 - 0.25\cos\left(\frac{\pi n}{150} + \frac{3\pi}{3}\right) \end{aligned}\right\} \quad n = 0, 1, \ldots, 399.$$

The overall signal $s(n)$ is shown in Fig. 1(a). The estimated frequency tracks are shown in Fig. 1(b). The amplitudes, obtained by solving a set of linear equations in the least-squares sense, are shown in Fig. 1(c). In solving for the frequencies, we found that (13) and (14) led to numerical difficulties even in the presence of small amounts of noise. This is because the argument of the inverse cosine function did not always have magnitude less than unity. On the other hand, frequency estimates obtained by rooting a polynomial were found to be more robust. Hence this approach was used.

We next applied this algorithm to speech data corresponding to the phoneme /oo/ (16 kHz sampling frequency). The data contained four formant frequencies around 500 Hz, 1200 Hz, 2300 Hz, and 3100 Hz., respectively. Subsequent low-pass filtering eliminated the third and the fourth formant frequencies. Fig. 2(a) shows the filtered speech signal. We used a model order $m = 8$ (four real sinusoids) on this data. Fig. 2(b) shows the first two formant frequency tracks, after smoothing by an eleven-point median filter. In Fig. 2(c) the corresponding least-squares estimates of the amplitude envelopes are shown, which used the median-filtered frequencies. Even though the filtered signal has only two dominant components, a model order of

---

[9] We thank Dr. Shubha Kadambe of A. I. duPont Institute, DE, for supplying us the speech data used in our simulations.

$m = 4$ (two real sinusoids) yielded poor estimates of the formant frequencies. Increasing the order to $m = 6$ improved the estimates noticeably; the estimates were, however, still noisy. A model order $m = 8$ was the smallest order required to yield reasonable results for this example.

7. CONCLUSIONS

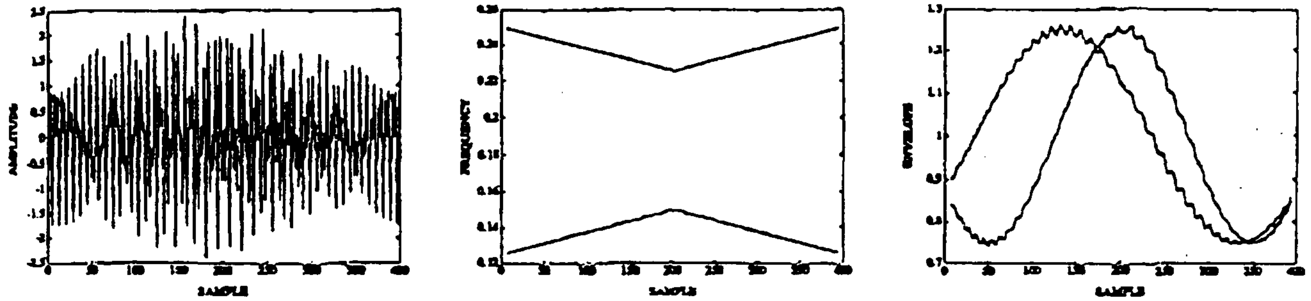
We have extended the Kaiser–Teager method to multicomponent signals. Even though this approach appears to be different from traditional frequency estimation algorithms, it bears a close resemblance to the Prony's method in that it requires polynomial rooting to estimate the component frequencies, particularly if m is larger than four (for complex signals) or eight (for real signals). It seems hard to escape from the clutches of Prony's method. There is scope for improving the method's performance in the presence of noise by choosing the $H_k(e^{j\omega})$ appropriately.

[4] I. H. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. ASSP-37, pp. 862–871, 1989.

[5] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE ICASSP-90*, (Albuquerque, NM), Apr. 1990.

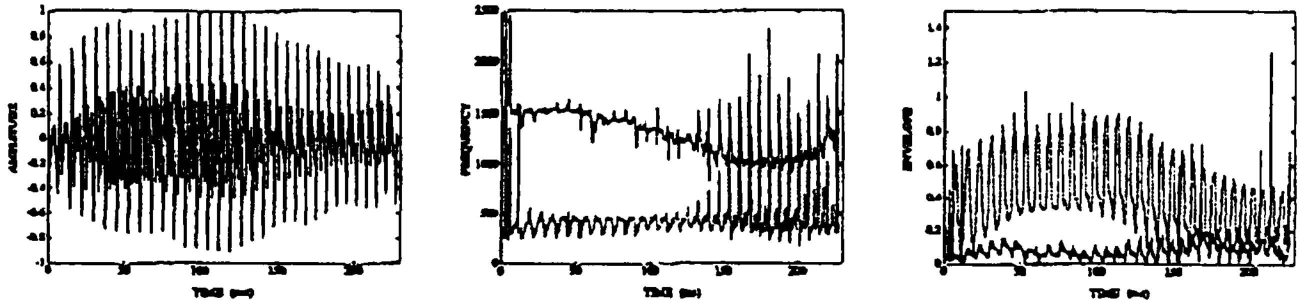
[6] J. F. Kaiser, "On Teager's energy algorithm and its generalization to continuous signals," in *Proc. IEEE DSP Workshop*, (New Paltz, NY), Sep. 1990.

[7] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators," in *Proc. IEEE ICASSP-92*, (San Francisco, CA), Mar. 1992.



(a) (b) (c)

Fig. 1 (a) Two-component AM-FM signal. (b) Estimated frequency tracks, smoothed by an 11-point median filter. (c) Estimated amplitude envelopes obtained via least-squares solution that utilized the median-filtered frequencies.



(a) (b) (c)

Fig. 2 (a) Signal from speech vowel /oo/, filtered to retain only the first two formant frequencies. (b) Estimated frequency tracks of the first and second formant frequencies, smoothed by an 11-point median filter. Assumed model order: $m = 8$. (c) Estimated amplitude envelopes obtained via least-squares solution that utilized the median-filtered frequencies.

REFERENCES

[1] R. J. MacAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," *IEEE Trans. Sig. Process.*, vol. ASSP-34, pp. 744–754, 1986.

[2] M. P. Cooke, "An explicit time–frequency characterization of synchrony in an auditory model," *Computer Speech and Language*, vol. 6, pp. 153–173, 1992.

[3] J. B. Y. Tsui, *Microwave receivers with electronic warfare applications*. New York, NY: John Wiley & Sons, 1986.

[8] R. Kumaresan and A. G. Sadasiv, "An invariance property of the determinant of a matrix whose elements are a sum of sinusoids and its application," *Archiv Elektrotechnik und Uebertragungstechnik*, 1992. Accepted for publication.

[9] F. B. Hildebrand, *Introduction to Numerical Analysis*, p. 458. New York: McGraw-Hill, second ed., 1974.

[10] R. A. Horn and C. A. Johnson, *Matrix Analysis*, p. 29. Cambridge: Cambridge University Press, 1985.

Paper # 2

ON ACCURATELY TRACKING THE HARMONIC COMPONENTS' PARAMETERS IN VOICED-SPEECH SEGMENTS AND SUBSEQUENT MODELING BY A TRANSFER FUNCTION*

R. Kumaresan *C. S. Ramalingam* *A. G. Sadasiv*
Department of Electrical Engineering, University of Rhode Island, Kingston, RI 02881

ABSTRACT

We propose an improved method to model voiced speech signals. First, we describe a method to accurately model the signals using a linear combination of harmonically related sinewaves. The method fits a linear combination of sines and cosines whose frequencies are integer multiples of the unknown fundamental (pitch) frequency to the speech data in the least-square sense. The amplitudes of the sinewaves and the fundamental frequency are the unknowns and are determined simultaneously using the least-squares fit. Using our method, we show how one can obtain smoothly varying frequency and amplitude tracks for all the harmonics and thus model the speech signal parsimoniously. After obtaining the harmonic decomposition, we regard the time-varying amplitudes of the cosinusoidal and sinusoidal harmonic components as the real and imaginary parts of the complex-valued frequency responses of the slowly time-varying filter representing the vocal tract and glottal excitation pulse generator, in cascade. We then fit a sequence of all-pole/pole-zero models to the complex frequency response values.

1. INTRODUCTION

Voiced-segments constitute a significant portion of speech signals. In many applications, it is important to extract features such as the pitch frequency and the vocal tract transfer function from these segments accurately, even when the speech signal is corrupted by noise. Usually, short-time Fourier transform (STFT), linear prediction (LP) or cepstral methods are used to extract these features.

Voiced-speech signals are often modeled as the output of a slowly time-varying linear filter representing the vocal tract, excited by a quasi-periodic glottal pulse train. If the glottal pulse train were indeed exactly periodic, it can be represented by a Fourier series with the fundamental frequency corresponding to the pitch frequency, which is given by the reciprocal of the period of the pulse train. Since the pulse train is only quasi-periodic, the voiced speech waveform may be modeled by a sum of harmonically related sinewaves with slowly varying fundamental frequency, with arbitrary amplitudes and phases. Many authors, perhaps starting with Flanagan, have observed this feature and taken advantage of it. Recently, McAulay and Quatieri [1, 2, 3, 4] (see also the references therein) have carried out extensive work in modeling both voiced and unvoiced speech by using a linear combination of sinusoidal signals. They have applied it to speech coding, co-channel interference suppression, and time scaling of speech. The algorithms proposed in the above references rely primarily on the STFT (or some modification of it) to obtain the sinewave decomposition.

In this paper, our primary contributions are two-fold:

- We describe a method to accurately estimate the fundamental/pitch frequency and the amplitudes of the harmonically related sinewaves simultaneously, using a direct least-squares fit to the speech data. Such methods are well known to model-based spectral analysis practitioners but appears not to have been used in speech analysis. We apply this method to a speech segment over short, possibly overlapping windows. Unlike in [1], we do not assume that the analysis window be an integer multiple of the pitch period or use the STFT peaks to determine the parameters. Also, we do not employ the pitch synchronous analysis advocated in [5]. Using our method we show how one can obtain smoothly varying frequency and amplitude tracks for all the harmonics and thus model the speech signal parsimoniously. This method is in fact the maximum-likelihood method, if the background noise is white and the assumed signal model is valid. Therefore, if the speech signal is corrupted by noise, it may be advantageous to estimate the harmonic components first using our method and then use them as 'cleaned-up' data for further modeling of the vocal tract etc.
- After obtaining the harmonic decomposition, we regard the time-varying amplitudes of the cosinusoidal and sinusoidal harmonic components as the real and imaginary parts of the complex-valued frequency responses of the time-varying filter, representing the vocal tract and glottal excitation pulse generator, in cascade. We assume that this cascaded filter is slowly time-varying. We then fit a sequence of transfer function models to the complex frequency response values.

2. ACCURATELY ESTIMATING THE HARMONIC COMPONENTS

Let us assume that a block of N samples of x_n, $n = 0, \ldots, N-1$, is to be modeled by a signal s_n consisting of M harmonically related sinewaves with unknown amplitudes

---

THIS RESEARCH WAS SUPPORTED BY AN AFOSR CONTRACT # F49620-92-J-0378

1068-6393/92 $03.00 © 1992 IEEE

nd fundamental frequency ω_0. M is assumed known.

$$s_n = \sum_{k=1}^{M} A_k \cos(\omega_0 kn) + B_k \sin(\omega_0 kn) \ . \qquad (1)$$

We wish to minimize the sum of squared errors by choosing he unknown amplitudes and the frequency ω_0.

$$E = \sum_{n=0}^{N-1} (x_n - s_n)^2. \qquad (2)$$

In matrix-vector notation, let $\mathbf{x} = (x_0, x_1, \ldots, x_{N-1})^T$ and $\mathbf{s} = (s_0, s_1, \ldots, s_{N-1})^T$ denote the data and signal model vectors, respectively. Using the model given in (1) we can write the signal vector $\mathbf{s}$ as

$$\mathbf{s} = \mathbf{W}\mathbf{a}, \qquad (3)$$

where $\mathbf{a}$ is the $2M \times 1$ vector of unknown amplitudes $\mathbf{a} = (A_1, A_2, \ldots, A_M, B_1, B_2, \ldots, B_M)^T$ and $\mathbf{W}$ is an $N \times 2M$ matrix, whose (k,l)-th element is given by

$$\mathbf{W}_{k,l} = \begin{cases} \cos(kl\omega_0) & l = 1, 2, \ldots, M \\ \sin(k(l-M)\omega_0) & l = M+1, M+2, \ldots, 2M \end{cases}$$

for $k = 0, 1, \ldots, N-1$. Using this notation we can rewrite the error E as

$$E = \|\mathbf{x} - \mathbf{W}\mathbf{a}\|_2^2 \qquad (4)$$

Since both $\mathbf{W}$ and $\mathbf{a}$ are unknown, this problem is a bilinear least-squares problem. Such problems have been dealt with in numerical analysis and spectral analysis literature for the past 20 years [6] (equation 16.152). The standard trick is to assume that ω_0 is known and then solve the least-squares problem for the best amplitudes. For a given ω_0 the best amplitude $\mathbf{a}$ is given by

$$\mathbf{a} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{x} \ . \qquad (5)$$

Substituting this value of $\mathbf{a}$ back into the error expression in (4) gives

$$E = \mathbf{x}^T(\mathbf{I} - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T)\mathbf{x}, \qquad (6)$$

where we have used the fact that the projection matrix $(\mathbf{I} - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T)$ is idempotent. Note that E now depends explicitly on the unknown ω_0 only. This error can be minimized by a coarse search over ω_0 followed by a gradient descent procedure. We have derived explicit expressions for the gradient and the Hessian, which are given in [7]. Once the best ω_0 that minimizes E is found, the corresponding amplitudes can be obtained from (5).

We mention two closely related problems. In the case of co-channel interference suppression [2], that is separating speech signals from two different speakers that have been added together, the same modeling procedure as above can be used. Except, in this case, s_n in (1) will be modeled as a linear combination of two sets of harmonically related sinewaves with two different fundamental frequencies. Now the error E in (6) will have to be minimized over two independent fundamental frequencies. In other situations, such as modeling unvoiced segments, the modeling procedure can be modified, to least-square fit a linear combination of sinusoids which are not necessarily harmonically related, to the data. In this case it may be *necessary to filter the data* into two or more frequency bands to reduce the number of sinusoids needed for the fit in each band.

3. ANALYSIS OF VOICED-SEGMENTS USING PROPOSED METHOD

Fig. 1 shows 3500 samples of the phoneme /oo/. The data was sampled at 16 kHz. Before applying the above algorithm to the speech data we low-pass filtered the data to about 1000 Hz and down-sampled by 4, for the following reasons:

- The low-pass region from 0-1 kHz contains the major portion of the signal energy.
- In this region the number of harmonic components with significant energy is likely to be small i.e., of the order of 10 or less and down-sampling reduces the required computation.
- Often the pitch frequency ω_0 varies slowly with time. This causes the harmonic components (some integer multiple of ω_0) in the high frequency range (say, near 3000 Hz) to sweep rapidly in frequency. We wish to exclude such components in our modeling, because the model in (1) is less valid for such components.

Fig. 2 shows the magnitude of the Fourier transform of the entire signal prior to filtering and down-sampling.

Next, we applied the algorithm described in section 2 to estimate ω_0 on overlapping blocks of data. Fig. 3(a) shows the error E as a function of the possible candidate ω_0's for the initial part of speech data. To find the minimum we first performed a coarse search to get a good initial guess and then used a gradient descent procedure [7] to find the global minimum of E which gave the best ω_0 estimate. Fig. 3(a) also shows the effect of block size on E as it is changed from one pitch period (32 samples) to about three pitch periods (96 samples). Note that the valley with the global minimum gets deeper as the size of the block increases. However, as the block size is increased, the optimal ω_0 also increases, because, in this example, the pitch frequency is slightly increasing with time. Fig. 3(b) shows the value of E as a function of the number of assumed harmonic components M (M varying from 2 to 5) for a fixed block size of 64 samples. Observe that the location of the minimum does not change much when M is chosen greater than 2. This shows that the precise value of M may not be that critical while estimating ω_0. Also observe that the DFT magnitude of the data clearly shows four or five distinct peaks in the frequency region from 0 to 1000 Hz.

The above method for estimating ω_0 is applied to contiguous overlapping blocks of data. The block size was 64 samples. The overlap was 60 samples. Fig. 4 shows the pitch frequency track thus obtained and its multiples, as a function of time (the dotted and solid curves; dotted curves have been used for some harmonics for ease of visualization). Next, we also estimated the frequencies of the underlying sinusoidal components *without* assuming that the sinewaves are harmonically related. This was done by a least-squares